International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 08 | Auguest-2016 www.irjet.net p-ISSN: 2395-0072

Data Deduplication using Even or Odd Block (EOB) Checking Algorithm in Hybrid Cloud

Suganthi.M[1], Hemalatha.B[2]

Research Scholar, Depart. Of Computer Science, Chikkanna Government Arts College, Tamilnadu, India[1].

Assistant Professor, Depart. Of Computer Science, Chikkanna Government Arts College, Tamilnadu, India[2].

Abstract - *Data deduplication is one of important data compression techniques for eliminating duplicate copies of repeating data, and has been widely used in cloud storage to reduce the amount of storage space and save bandwidth. Data deduplication is a capacity optimization technology that is being used to dramatically improve storage efficiency. IT managers and executives face explosive data growth, driving up costs of storage for backup and disaster recovery (DR). For this reason, data deduplication is regarded as the next evolutionary step in backup technology and a "must-have" for organizations that wish to remain competitive by operating as efficiently as possible. Data deduplication maximizes storage utilization while allowing IT to retain more near line backup data for a longer time. This tremendously improves the efficiency of disk-based backup, changing the way data is protected. In general, data deduplication compares new data with existing data from previous backup or archiving jobs, and eliminates the redundancies. Advantages include improved storage efficiency and cost savings, as well as bandwidth minimization for less-expensive and faster offsite replication of backup data. We use Even or Odd Block (EOB) checking algorithm and metadata manager MM for data deduplication in hybrid cloud as it reduce the time instead of comparing all blocks in the file.*

Key Words: Data Deduplication, EOB Algorithm, Metadata Manager, Cloud Storage, Keyword File.

1. INTRODUCTION

Cloud systems [1, 2] can be used to enable data sharing capabilities and this can provide an abundant of benefits to the user. There is currently a push for IT organizations to increase their data sharing efforts. According to a survey by InformationWeek [3], nearly all organizations shared their data somehow with 74 % sharing their data with customers and 64 % sharing with suppliers. A fourth of the surveyed organizations consider data sharing a top priority. The benefits organizations can gain from data sharing is higher productivity. With multiple users from different organizations contributing to data in the Cloud, the time and cost will be much less compared to having to manually exchange data and hence creating a cluster of redundant and possibly out-of-date documents. With social networking services such as Face book, the benefits of sharing data are numerous [4] such as the ability to share photos, videos, information and events, create a sense of enhanced enjoyment in one's life and can enrich the lives of some people as they are amazed at how many people are interested in their life and well-being. For students and group-related projects, there has been a major importance for group collaborative tools [5]. Google Docs provides data sharing capabilities as groups of students or teams working on a project can share documents and can collaborate with each other effectively. This allows higher productivity compared to previous methods of continually sending updated versions of a document to members of the group via email attachments. Also in modern healthcare environments, healthcare providers are willing to store and share electronic medical records via the Cloud and hence remove the geographical dependence between healthcare provider and patient [6].

Some of major requirements of secure data sharing in the Cloud are as follows. Firstly the data owner should be able to specify a group of users that are allowed to view his or her data. Any member within the group should be able to gain access to the data anytime, anywhere without the data owner's intervention. No-one, other than the data owner and the members of the group, should gain access to the data, including the Cloud Service Provider. The data owner should be able to add new users to the group. The data owner should also be able to revoke access rights against any member of the group over his or her shared data. No member of the group should be allowed to revoke rights or join new user to the group.

In general, data deduplication compares new data with existing data from previous backup or archiving jobs, and eliminates the redundancies [7]. Advantages include improved storage efficiency and cost savings, as well as bandwidth minimization for less-expensive and faster offsite replication of backup data.

File-interface Deduplication System (FDS), offer flexible, high-performance data deduplication to enable organizations to overcome backup challenges, optimize capacity, reduce storage costs, and minimize WAN requirement.

Data deduplication works by comparing blocks of data or objects (files) in order to detect duplicates. Deduplication can take place at two levels — file and sub-file level. In some systems, only complete files are compared, which is called Single Instance Storage (SIS). This is not as efficient as sub-file deduplication, as entire files have to be stored again as a result of any minor modification to that file.

Even or Odd Block (EOB) checking algorithm used to provide sub-file or block-based deduplication. If user want to upload a file to cloud before uploading check it already exists or not. Data are broken into blocks and assigned an index number and also get keywords from metadata manager. The keywords of created file maintained separately for future purpose. First check the even blocks of the file based on that keywords .If any match occurs omit that file otherwise check it with odd block. At the time of checking odd block, any odd block matched that keywords, the file already exist so use the stored file otherwise save and upload that file. So EOB algorithm reduce the time unless check all data blocks. Once it is determined that a block of data already exists in the deduplication repository, the block is replaced with a Virtual Index Pointer linking the new sub-block to the existing block of data in the repository. If the sub-block of data is unique, it is stored in the deduplication repository and a virtual index is stored in memory for fast comparison.

2. RELATED WORK

Many systems have been developed to provide deduplication in the cloud storage but that is not suitable.

1) File-level de-duplication: It is commonly known as single-instance storage, file-level data de-duplication compares a file that has to be archived or backup that has already been stored by checking all its attributes against the index. The index is updated and stored only if the file is unique, if not than only a pointer to the existing file that is stored references. Only the single instance of file is saved in the result and relevant copies are replaced by"stub" which points to the original file.

2) Block-level de-duplication: Block-level data deduplication [8] operates on the basis of sub-file level. As the name implies, that the file is being broken into segments blocks or chunks that will be examined for previously stored information vs redundancy. The popular approach to determine redundant data is by assigning identifier to chunk of data, by using hash algorithm for example it generates a unique ID to that particular block. The particular unique Id will be compared with the central index. In case the ID is already present, then it represents that before only the data is processed and stored before .Due to this only a pointer reference is saved in the location of previously stored data. If the ID is new and does not exist, then that block is unique. After storing the unique chunk the unique ID is updated into the Index. There is change in size of chunk as per the vendor. Some will have fixed block sizes, while some others use variable block sizes.

In the existing system deduplication method applied to each block-block or byte-byte so it is time consuming for huge data in the cloud storage. Low level of data integrity.

3. PROPOSED SYSTEM:

To overcome the problem of existing system we apply Even or Odd Block (EOB) checking algorithm for finding duplicates in short time. It is efficient and time consuming. In this system we apply duplication check in the client system, server based security mechanism.

Network deduplication is used to reduce the number of byte that must be transferred between endpoints, which can reduce the amount of bandwidth required and also transmission cost.

Metadata Manager (MM)

MM is the component responsible for storing metadata, which include file table, pointer table, and keyword table.MM maintains a linked list and a small database in order to keep track of file ownerships, file composition and avoid the storage of multiple copies of the same data segments.

The tables used by MM are structured as follows:

File table: The file table contains the file id, file name,user id and the id of the first data block.

Pointer table: The pointer table contains the block id and the id of the block stored at the cloud storage provider.

Keyword table: The Keyword table contains the important keys of the stored file in the cloud storage server.
In addition to the access control mechanism performed by the server, when users ask to retrieve a file, MM further checks if the requesting user is authorized to retrieve that file.

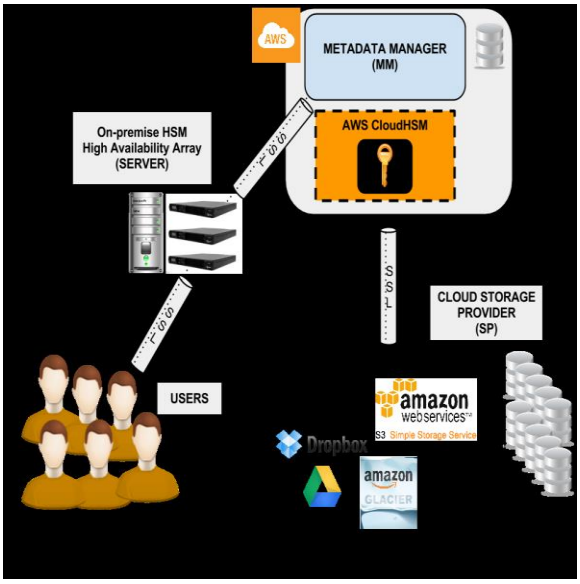

Fig 1: A realistic example of cloudDedup

This way, MM makes sure that the user is not trying to access someone else's data. This operation can be considered as an additional access control mechanism, since an access control mechanism already takes place at the server. Another important role of MM is to communicate with cloud storage provider (SP) in order to actually store and retrieve the data blocks and get a pointer to the actual location of each data block.

Even or Odd Block Checking (EOB) Algorithm:

Even or Odd Block (EOB) checking algorithm used to provide sub-file or block-based deduplication. If user want to upload a file first send file detail like user id, file id, keyword of the file to MM.MM check whether the file already stored or not. If any file matched or already stored that files can be checked again using EOB algorithm by extracting that files from storage server.
Data are broken into blocks and assigned an index number. The keywords of created file maintained separately for future purpose. First check, the even blocks of the new file with existing file. If any match occurs omit that file otherwise check it with odd block.

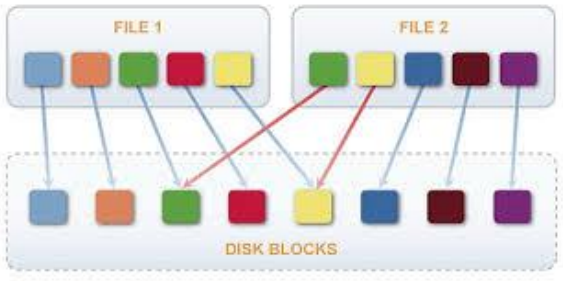

Fig 2: Comparing File Blocks

At the time of checking odd block, any odd block matched with stored file,the file already exist so use the stored file otherwise save and upload new file. So EOB algorithm reduce the time unless check all data blocks. Once it is determined that a block of data already exists in the deduplication repository, the block is replaced with a Virtual Index Pointer linking the new sub-block to the existing block of data in the repository. If the sub-block of data is unique, it is stored in the deduplication repository and a virtual index is stored in memory for fast comparison.

EOB Algorithm:

$File is a passed array containing four keys->File Name n, User_id uid, Keyword k, Date d
**$**EOB str=$File ('File Name');
**$**EOB str=$File ('User_id');
**$**EOB str=$File ('Keyword');
**$**EOB str=$File ('Date');
**$**Unique File id=$File ('File005',$EOBstr);
//if file already stored again check by current file using EOB
Clear
Echo 'split file into block and assign index number i'
Read i
Let r=$ i%2
If($r –eq 0)
Then
Echo$ i is Even //check all even block
Let File1 Block b1=File2 Block b2
If(b1=b2)
Then
Echo$ 'File Already exist'
Else
Echo$ ' Go to odd block'
end
Echo $ i Odd //check all odd block
If(b1=b2)
Echo$ 'File Already exist'

Else
Echo$ 'strore and upload file'
End

The simple idea behind deduplication is to store duplicate data (either file or block) only once. Therefore, if a user wants to upload a file which is already stored, the cloud provider will add the user to the owner of the file (block).

4. FILE RETRIEVAL

During the retrieval procedure, a user asks to download a file from the system. Fig 3 describes a scenario in which a user Uj wants to download the file F1.

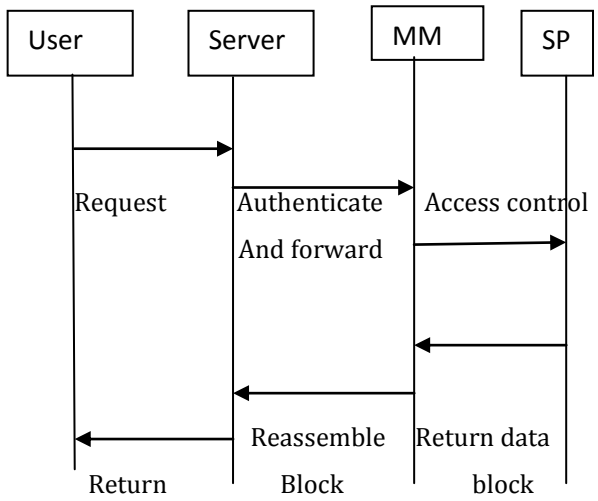

Fig 3: File Retrieval Protocol

MM-Metadata Manager **SP**-Storage Provider

USER: User Uj sends a retrieval request to the server in order to retrieve file F1. The request is composed by the user's id I DUj , the file identifier Fid1 and his certificate.
SERVER: The server receives the request, authenticates Uj and if the authentication does not fail, the server forwards the request to MM.
MM: MM receives the request from the server and analyzes it in order to check if Uj is authorized to access Fid1 (Uj is the owner of the file). If the user is authorized, MM looks up the file identifier in the file table in order to get the pointer to the first block of the file. Then, MM visits the linked list in order to retrieve all the blocks that compose the file. For each of these blocks, MM retrieves the pointer from the pointer table and sends a request to SP.

5. IMPLEMENTATION:

To implement a new efficient algorithm for avoid duplicate data in the hybrid cloud storage. A client program is used to model the data users to carry out the file upload process and create a Keyword File. Implementation focus on comparing the overhead induced by following step, including authorization, keyword generation and maintain file detail like filename, file id, user id.
Breakdown the upload process four steps 1.MM 2.EOB 3.Duplicate check 4.Transfer. For each step, record start and end time of it and therefore obtain the breakdown of the total time spent.
To implement the communication between entities based on HTTP, using GNU Libmicrohttp[9] and libcurl[10]. Thus, users can issue HTTP Post request to the server. Fig 4 shows the deduplication ratio in time (sec) based on overall function performed from file creation to upload. In between several processes happened. Initially MM check the file with stored file detail then apply EOB for finding duplication and finally upload the file.

Deduplication Ratio

To evaluate the effect of the deduplication ratio, user prepares two unique data sets, each of which consists of 50 100MB files. User first upload the first set as an initial upload. For the second upload, we pick a portion of 50 files, according to the given deduplication ratio, from the initial set as duplicate files and remaining files from the second set as unique files.

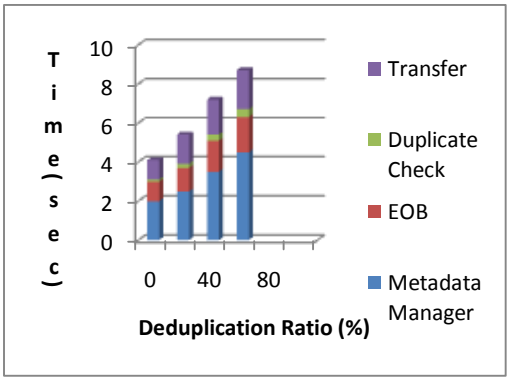

Chart -1: Time breakdown for different duplication Ratio

6. EVALUATION

In this section we evaluate the overhead introduced by our system in terms of storage space and computational complexity. In order to refer to a real scenario, we use the same parameters of [11], but our calculations hold true for Other scenarios.

A. Storage Space

We took into account a scenario in which there are 857 file systems. The mean number of files per file system is 225K and the mean size of a file is 318K, resulting in about 57T of data.

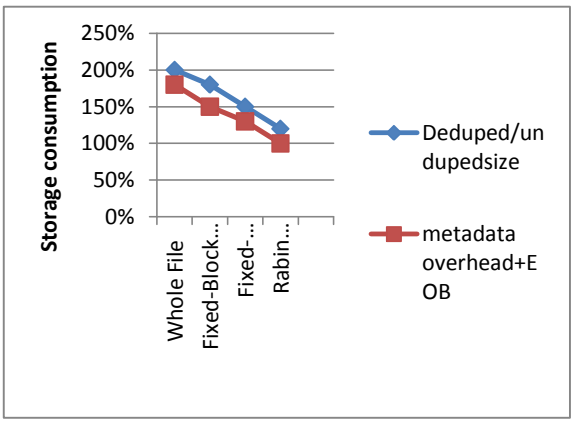

Chart -2: Metadata management with EOB

Number of Stored Files

To evaluate the effect of number of stored files in the system, user upload 10000 10MB unique files to the system and record the breakdown for every file upload. From Despite of the possibility of a linear search, the time taken in duplicate check remains stable due to the low collision probability.

7. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have reviewed literature on ways to remove duplicate copies of data using Even or Odd Block Checking Algorithm. So we save space and transmission cost. Future research direction would be to give the data owner physical access control over his data. Instead of accountability, the data owner can create a set of access control rules on his data and send the data along with the access control policy. In this way, any member with access to the data can only use the data in such a way that abides by the access control policy. If a member attempts to make illegal copies of the data, the access control policy should "lock" the data to prevent the member from doing so.
Also, since data stored in the Cloud are usually stored and replicated in different geographical locations around the world, it is crucial that the legal jurisdictions are honored and followed. A potential research direction would be to find ways to store and process data in a way that does not breach the privacy and security laws of the region.

REFERENCES

[1] Mell P, Grance T (2012) The NIST definition of cloud computing. NIST Spec Publ 800:145.National Institute of Standards and Technology, U.S. Department of Commerce. Source: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf. Accessed on Oct 2012

[2] Wikipedia definition of Cloud computing (2012).Source: http://en.wikipedia.org/wiki/Cloud_computing. Accessed on Oct 2012

[3] HealeyM(2010) Why IT needs to push data sharing efforts. InformationWeek.Source:Http://www.informationweek.com/services/integration/why-it-needs-to-push-data-sharing-effort/225700544. Accessed on Oct 2012

[4] Gellin A (2012) Facebook's benefits make it worthwhile. Buffalo News.

[5] Riley DA (2010) Using google wave and docs for group collaboration. Library Hi Tech News.

[6] Wu R (2012) Secure sharing of electronic medical records in cloud computing. Arizona State University, ProQuest Dissertations and Theses

[7] Pandey S,VoorsluysW, Niu S,Khandoker A, BuyyaR(2012) An autonomic cloud environment for hosting ECG data analysis services. Future Gener Comput Syst 28(1):147–154

[8] For more information, visit www.falconstor.com or contact your local FalconStor representative.

[9] Advanced Data Deduplication Techniques And Their Appl Ication Dirk Meister

[10] Dutch T Meyer and William J Bolosky. A study of practical deduplication. ACM Transactions on Storage (TOS), 7(4):14, 2012.

[11] Google Drive. http://drive.google.com/.